Statistical Learning

Statistics 640 • Fall 2012 • Mondays & Wednesdays 2:30 - 3:45pm

Instructor:
Genevera Allen
Office: Duncan Hall 2098.
Office Hour: Wednesday 4 - 5pm
Email: gallen@rice.edu.

Teaching Assistant:
Yue Hu
Office: Duncan Hall 1038.
Office Hours: Tuesdays 12:30 - 2:30pm
Email: yh6@rice.edu.

Time & Location:
Time: Mondays & Wednesdays 2:30 - 3:45pm. Location: Mechanical Laboratory 251.

Course Description: This course is a survey of statistical learning methods. The course will be loosely divided into two sections covering optimization-based methods and algorithmic-based methods for both supervised and unsupervised learning. Topics covered include penalized regression and classification, support vector machines, kernel methods, model selection, matrix factorizations and completion, graphical models, clustering, boosting and ensemble learning.

Course Objectives & Outcomes: Students will learn how and when to apply statistical learning techniques, their comparative strengths and weaknesses, and how to critically evaluate the performance of learning algorithms. Students completing this course should be able to (i) apply basic statistical learning methods to build predictive models, (ii) properly tune and select statistical learning models, (iii) correctly assess model fit and error, and (iv) build an ensemble of learning algorithms.

Course Tracks: This course will be offered as two tracks, an *Applied Track*, and a *Theory and Methods Track*. The two tracks will be assigned different homeworks and be given a different mid-term exam. The class will jointly participate in the data mining competition. Students may choose to focus on one track or both and must designate this when the first homework assignment is due.

Prerequisites: (Please disregard course prerequisites listed in the course catalog). Required: Linear Algebra & Scientific Programming in a language such as `R`, `Matlab`, or `Python`. Recommended: Linear Regression. Please speak to the instructor if you have any concerns regarding prerequisites. The instructor will sign prerequisite waiver forms for any interested student.

Textbook: Elements of Statistical Learning by Hastie, Tibshirani & Friedman. Available online at `http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf` and book webpage at `http://www-stat.stanford.edu/~tibs/ElemStatLearn`.

Course Webpage: `http://www.stat.rice.edu/~gallen/stat640.html`. Please check the course webpage frequently for announcements and homework assignments.

Grading Policy:

Homeworks	15%
Mid-Term Exam	35%
Competition & Reports	45%
Class Participation	5%

Registrar Deadlines:

Friday, August 31	Add Deadline
Friday, October 5	Drop Deadline

Homeworks: There will be three homework assignments consisting of mathematical proofs and derivations (Theory & Methods Track) and/or data analysis (Applied Track). Approximate dates:

	Homework Assigned	Homework Due
Homework I	September 5	September 26
Homework II	October 1	October 24
Homework III	November 5	November 21

A hard copy of homeworks are due at 5:00pm to the TA in Duncan Hall on the due date. Late homeworks will NOT be accepted. Homeworks may be discussed with classmates but must be written and submitted individually.

Mid-Term Exam: There will be an in-class mid-term exam on **Wednesday, October 31**. The exam will be open book and open notes. As an incentive, the team at the top of the leaderboard of the data mining competition will receive five extra points on the mid-term exam.

Data Mining Competition:
As part of the class, students will compete against each other in a data mining contest. The competition will begin on Wednesday, September 5 and can be completed in teams of two people. Grades will be based upon two progress reports and a final report (one per team) as well as the contest results. The winning team in terms of prediction accuracy and the team with the most innovative solution will automatically receive the highest grades. Further details about the contest along with specific grading criteria will be given in a separate document and discussed in class.

Contest Opens	12:00am September 5
Progress Report I	5:00pm October 3
Progress Report II	5:00pm November 5
Contest Closes	11:59pm November 27
In-Class Contest Presentations	November 28
Final Report Due	5:00pm November 30

Tentative Schedule:

Part I: Optimization-Based Methods		
	August 20	Introduction
	August 22	Least Squares & Ridge
	August 27	Ridge, PCA & PLS Regression
	August 29	Review: Matrix Algebra & Optimization
	September 5	Sparse Regression I
	September 10	Sparse Regression II
	September 12	Linear Discriminant Analysis
	September 17	LDA & Logistic Regression
	September 19	Sparse Classification
	September 24	Support Vector Machines I
	September 26	SVMs II
	October 1	Kernel Methods
	October 3	Model Selection I
	October 10	Model Selection II
	October 15	Matrix Factorizations I
	October 17	Matrix Factorizations II
	October 22	Matrix Factorizations & Completion
	October 24	Graphical Models I
	October 29	Graphical Models II
Midterm Exam	October 31	
Part II: Algorithmic-Based Methods		
	November 5	Clustering I
	November 7	Clustering II
	November 12	CART
	November 14	Ensemble Learning
	November 19	Boosting I
	November 21	Boosting II
	November 26	Random Forests
	November 28	Recap: Data Mining Competition
Contest Closes	November 27	
Final Report Due	November 30	

Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.